Fine-tuning language models for emotion recognition in Lithuanian texts using neural machine translation of training datasets

Mindaugas Petkevičius[a], Daiva Vitkutė-Adžgauskienė[a]

[a] *Vytautas Magnus University, K. Donelaičio g. 58, Kaunas, 44248, Lithuania*

Abstract

Lithuanian language is a complex and rich language with a unique grammar structure, making it an interesting choice for natural language processing (NLP) tasks such as emotion detection. This study provides helpful insights into the emotional nuances of Lithuanian texts by utilizing a translated and augmented emotion dataset. We present a methodology that leverages translated datasets for emotion recognition and augmentation approaches to improve the performance of emotion identification models in low-resource languages. We compared the outcomes of transformer-based language models, such as RoBERTa, LaBSE, and LitLat BERT on the translated and augmented data. Our results demonstrated that LitLat BERT, which is primarily trained on Lithuanian texts, showed the most significant improvement in performance when data augmentation was applied. We conclude that LitLat BERT could be the preferred choice for emotion recognition tasks for Lithuanian language due to its specialized training and enhanced adaptability when provided with diverse and augmented data. This study provides valuable insights into the challenges and potential solutions for emotion identification tasks in morphology-rich languages, like Lithuanian language.

Keywords

NLP, emotion detection, transformer-based models, BERT, translation, augmentation, Lithuanian language

1. Introduction

Emotion recognition in texts is an important natural language processing (NLP) task and has wide-ranging applications, including sentiment analysis, customer support, mental health monitoring and others. With the rapid development of deep learning techniques, significant advancements have been made in emotion recognition for high-resource languages, such as English, Chinese and Spanish. However, research in low-resource languages, like Lithuanian language, remains limited due to the scarcity of annotated datasets and the lack of pre-trained models.

In this paper, we propose a methodology for improving emotion recognition performance in low-resource and morphology-rich languages, using translated datasets and augmentation approaches for the Lithuanian language.

In order to reach our goal, we perform the following tasks: related work analysis (Section 2), dataset analysis (Section 3), methodology for translating and augmenting datasets, as well as fine-tuning transformer-based language models (Section 4), experimental evaluation of different fine-tuned models (Section 5), conclusions and future plans (Section 6).

28th International Conference Information Society and University Studies - IVUS 2023
EMAIL: mindaugas.petkevicius@vdu.lt (M. Petkevičius); daiva.vitkute@vdu.lt (D. Vitkutė-Adžgauskienė)
ORCID: 0000-0002-1120-4848 (A. 1); 0000-0001-7923-1087 (D. Vitkutė-Adžgauskienė)

2. Related Works

Research in emotion recognition for low-resource languages has been an area of growing interest in recent years [1]. Several studies have proposed different approaches to tackle this problem, such as cross-lingual transfer learning [2], data augmentation technique [3], and leveraging unsupervised methods [4].

Also there have been studies, where emotion detection datasets across 19 languages were trained on a multilingual emotion prediction model, XLM-EMO, for social media data, which showed competitive performance in a zero-shot setting and is particularly useful for low-resource languages [5].

In another study, researchers addressed the shortage of annotated gold standard resources for emotion mining by presenting a multilingual emotion dataset of tweets in English and Spanish, labeled with one of seven emotions and proposed a machine learning approach for automatically detecting emotions in tweets for both languages [6].

Also, researchers conducted experiments on sentiment and emotion recognition for English and Polish texts in the context of a chatbot, creating a parallel corpus named CORTEX and employing various classifiers such as Support Vector Machines, fastText, and BERT, with BERT-based models yielding the highest accuracy and F1-scores, although results for Polish were slightly inferior to those for English [7].

However, these methods are often problematic due to the need for parallel data, noise adding, or relying on uncontrolled training goals that may not match the target task.

3. Dataset

We selected the CARER (Contextualized Affect Representations for Emotion Recognition) [8] dataset, a widely-used and well-annotated dataset for emotion recognition tasks for our research in this study. The dataset contains textual data, labeled with six different emotions: *sadness*, *joy*, *love*, *anger*, *fear*, and *surprise*. In this section, we provide a brief overview of the dataset and discuss the reasons for choosing it for our emotion recognition task for the Lithuanian language.

Table 1
Dataset text split

Type	Documents
Training	16 000
Validation	2 000
Testing	2 000
Total	20 000

This dataset is a collection of 20,000 English social media text samples, with each sample annotated with one of the six emotions. The dataset is split in the following way: 16,000 samples for training, 2,000 samples for validation, and 2,000 samples for testing. The data is diverse, containing a variety of sentence structures and extensive vocabulary, allowing the models to learn the nuances of different emotions in textual data. Label distribution is presented in Figure 1 and text length distribution is presented in Figure 2.

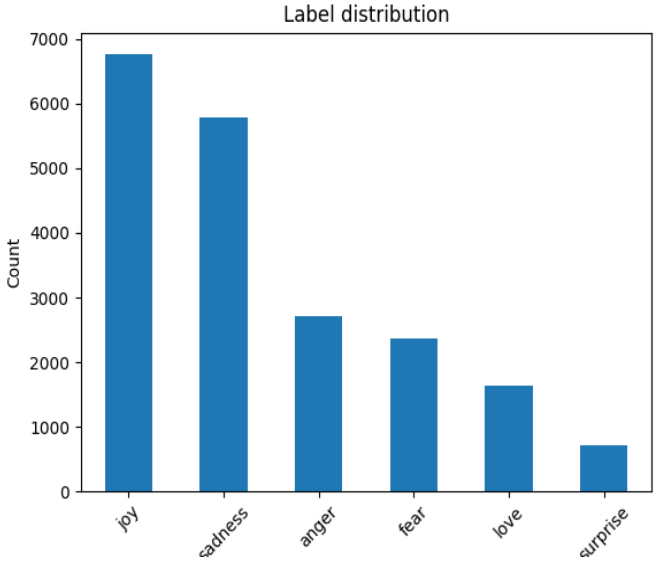

Figure 1: Label distribution

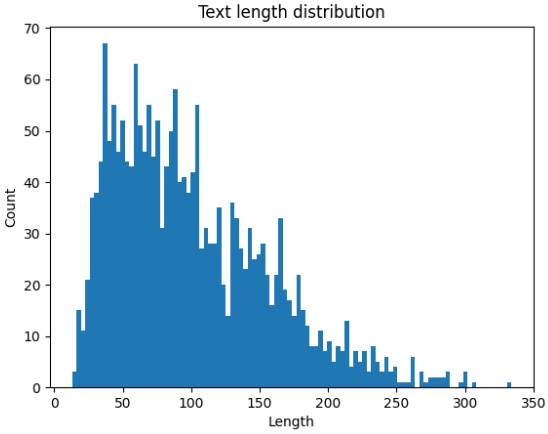

Figure 2: Text length distribution

There were several reasons for selecting this dataset for our study:

1. High-quality annotations: accurate and well-annotated labels for model training and evaluation.
2. Diversity of emotions: wide range of emotions for learning linguistic patterns and improved emotion recognition.
3. Transfer learning suitability: English dataset enables transfer learning for low-resource Lithuanian tasks using translations.
4. Compatibility with models/frameworks: Widely used dataset, compatible with pre-trained models, simplifies incorporation and focuses on emotion recognition challenges.

Summarizing, the dataset was chosen due to its high-quality annotations, diversity of emotions, suitability for transfer learning, and compatibility with existing models and frameworks. By leveraging this dataset in conjunction with language-specific models and translation techniques, we can effectively address the challenges associated with emotion identification in low-resource languages such as Lithuanian.

4. Methodology

The methodology part covers the methods applied in this research: (1) translation of the dataset using neural machine translation models; (2) augmentation process for enriching and balancing the dataset; (3) fine-tuning transformer-based models for emotion classification task.

4.1. Translation of the dataset

For the emotion recognition task, the choice of translation service plays a crucial role in generating high-quality translated datasets. To increase the linguistic diversity of our dataset, we translated these datasets using two state-of-the-art machine translation models:

1. **Opus-MT** [9]: An open-source neural machine translation model trained on a variety of multilingual parallel corpora.
2. **facebook/m2m100_418M** [10] A many-to-many multilingual translation model developed by Facebook AI [11]. It is designed to support translation between 100 languages and is trained on a large-scale parallel corpus.

However, depending on specific requirements and constraints of certain projects, DeepL [12] or Google Translate [13] could also be viable options.

By combining and translating these datasets, we aimed to create a rich, diverse dataset that covers a wide range of emotions, contexts and social texts.

When choosing a translation service, it is important to think about the quality of the translation, the number of domains it covers, and how well it fits the task at hand. In some situations, it might be helpful to improve the quality of the translated information by using a combined approach, merging translations from more than one service.

After combining the translations, generated by both models, we obtained a dataset with the following distribution of emotion labels, as displayed on Figure 3.

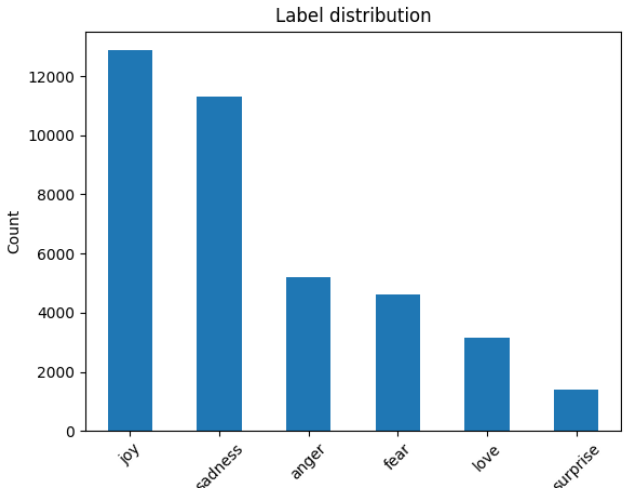

Figure 3: Labels of a combined dataset

It was observed that certain emotion classes, specifically *love* and *surprise*, have a significantly lower number of samples in the dataset. To address this imbalance and improve the performance of our

models, we decided to augment the dataset by applying synonym replacement to the underrepresented classes. Synonym replacement has showed good results for dataset augmentation [14].

4.2. Augmentation

In our study, we employed data augmentation techniques to enhance the performance of the emotion recognition models and improve their robustness to different translations and imbalanced emotion distributions. For this purpose, we used synonym replacement as our data augmentation method. In this section, we describe the process of synonym replacement and its importance in the context of the Lithuanian language.

Synonym replacement is accomplished by replacing words in a given text with their synonyms in order to create new and slightly different variations of the original text while preserving its meaning and sentiment. The process of synonym replacement is shown in Figure 4.

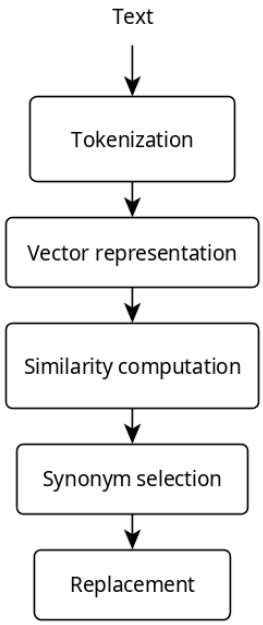

Figure 4: Process of a synonym replacement

We used the following algorithm to implement synonym replacement in our study:

1. Tokenization: Tokenize the Lithuanian text samples using a spaCy [15] library tokenizer.
2. Vector representation: Represent each word token using the GloVe [16] word vectors provided by the spaCy library. These vectors capture the semantic word meaning in a high-dimensional space.
3. Similarity computation: For each word token, compute the cosine similarity between its vector representation and the vectors of other words in the vocabulary to identify the most semantically similar words like synonyms.
4. Synonym selection: Select the most similar words, matching the same part-of-speech (POS) and morphology, as the original word token. This is crucial for preserving the grammatical correctness of the text, especially in a morphologically rich language like Lithuanian.
5. Replacement: Randomly replace a subset of word tokens in the text with their selected synonyms, ensuring that the replacements do not alter the text's overall meaning or sentiment.

Synonym replacement increases the performance of the emotion recognition model for Lithuanian by addressing morphological richness, promoting diverse training data, and handling imbalanced datasets.

It preserves grammatical correctness and overall sentiment, exposes models to various linguistic patterns, and generates additional samples for underrepresented emotion classes.

After analyzing the dataset, we decided to add additional augmented texts for the labels *love* and *surprise*, which had the fewest occurrences. The updated distribution of emotion labels in the dataset is presented in Figure 5.

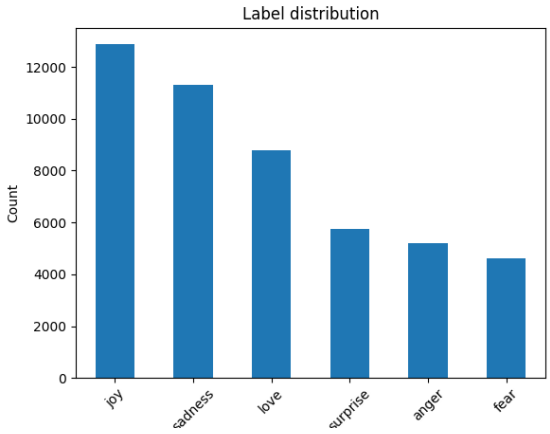

Figure 5: Label distribution after augmentation

This balanced distribution of emotion labels enables the models to learn better representations of different emotions in the text, and potentially improves their overall performance in the emotion identification task. Final version of the dataset is shown in Table 2.

Table 2
Final version of the dataset: label distribution

Label	Count
joy	12805
sadness	11213
love	8260
surprise	5422
anger	5173
fear	4564
Total	47437

4.3. Training and fine-tuning language models

We chose three transformer-based models for our experiments for investigating the impact of translation quality and data augmentation on emotion recognition tasks in Lithuanian:

1. *RoBERTa* [17] (xlm-roberta-base): A multilingual version of RoBERTa pretrained on various languages, offering state-of-the-art performance in numerous NLP tasks.
2. *LaBSE* [18] (sentence-transformers): A sentence-level transformer architecture designed for cross-lingual tasks, generating fixed-size representations for multilingual text data.
3. *Litlat BERT* [19] (litlat-bert): A fine-tuned BERT [20] model tailored to Lithuanian and Latvian languages, pretrained on a large corpus of Lithuanian and Latvian text data.

By comparing these diverse models, we aimed to identify the most efficient approach for emotion recognition for the Lithuanian language, and evaluate the impact of translation quality and data augmentation techniques on their performance.

5. Experiments and results

Experiments were carried out in a series of tasks:

1. Firstly, the quality of translations was assessed.
2. Secondly, base model for English language was fine-tuned using original English dataset.
3. Thirdly, the obtained word embedding models were evaluated using the adapted intrinsic evaluation benchmarks.
4. Finally, the resulting data was examined in order to determine the effect of different hyperparameters on benchmark evaluation results.

5.1. Assessing quality of the translations

In our study, we used two translation models to generate Lithuanian versions of the dataset. Due to the morphological richness of Lithuanian, achieving consistent translations can be challenging. Translation models may struggle to generate accurate translations, which can impact emotion identification performance.

However, as long as the main sentiment remains consistent, this impact is expected to be minimal. Training models on diverse translation styles can improve their robustness and performance in real-world scenarios [21].

To assess translation quality, we calculated the **BLEU** (bilingual evaluation understudy) 0.2337 and **WER** (word error rate) 0.3349 scores for the translations. These scores indicate significant differences between the two models, attributable to the complex nature of the Lithuanian language. Despite the differences, the main idea is generally preserved, enabling emotion identification models to learn from the dataset. This highlights the importance of translation quality in low-resource language tasks and the need for more robust models to handle translation inconsistencies.

5.2. Base model trained on original English dataset

In order to establish a baseline for comparison, we trained a base model on the original English dataset. This experiment allowed us to evaluate the performance of the emotion identification models when trained on a high-resource language, like the English language, and compare the results with those obtained for the Lithuanian translations.

We selected the *xlm-roberta-base* model for this experiment, as it has demonstrated strong performance in various NLP tasks [22], including emotion identification. The results obtained from training the *xlm-roberta-base* model on the original English dataset are presented in Table 3:

Table 3
Base model training results

Model	Accuracy	F1	Precision	Recall
XLM-RoBERTa (xlm-roberta-base)	0.9305	0.9312	0.9330	0.9305

As expected, the *xlm-roberta-base* model achieved a high level of performance when trained on the English dataset, with an accuracy of **0.9305** and an F1 score of **0.9313**. These results serve as a benchmark for assessing the performance of the emotion identification models trained on the translated Lithuanian datasets, enabling us to better understand the challenges and potential improvements associated with semantically rich and low-resource languages.

5.3. Fine-tuning models for emotion detection

In the second training experiment, we aimed to evaluate the performance of the emotion identification models when exposed to a diverse range of translation styles and potential errors. For this purpose, we combined the datasets generated using both translation models. This approach provided the models with a more varied and potentially challenging dataset, allowing us to assess the model's adaptability and robustness in the face of translation inconsistencies. Results are provided in Table 4.

Table 4
Performance of the emotion identification models

Model	Accuracy	F1	Precision	Recall
XLM-RoBERTa (xlm-roberta-base)	0.800	0.7995	0.8002	0.8002
LaBSE (sentence-transformers)	**0.803**	**0.800**	**0.8010**	**0.8036**
LitLat BERT (litlat-bert)	0.795	0.7935	0.7925	0.7953

Comparing the results of the experiments, we observed that when trained on the original English dataset, the *xlm-roberta-base* model achieved the highest accuracy, precision, recall, and F1 score. However, when using translated dataset, the performance of *xlm-roberta-base* decreased but remained competitive. In this scenario, LaBSE showed slightly better accuracy and F1 score compared to *lm-roberta-base*, while LitLat BERT (litlat-bert) had marginally lower results.

This comparison demonstrates that translation quality and the choice of translation model can impact the performance of emotion identification models in low-resource languages like Lithuanian. Moreover, they emphasize the need for developing more robust models that can effectively handle translation inconsistencies and variations in the text.

5.4. Fine-tuning models for emotion detection with augmented dataset

In the third training experiment, we aimed to evaluate the performance of the emotion identification models when exposed to a diverse range of translation styles, potential errors, and augmented data. For this purpose, we used the translated dataset, and, additionally, applied data augmentation techniques. This approach provided models with a more varied and potentially challenging dataset, allowing us to assess the models' adaptability and robustness in the face of translation inconsistencies and the additional variations introduced by data augmentation. Table 5 shows the results.

Table 5
Performance of the emotion identification models with augmentation

Model	Accuracy	F1	Precision	Recall
XLM-RoBERTa (xlm-roberta-base)	0.810	0.8097	0.8100	0.8101
LaBSE (sentence-transformers)	0.781	0.7795	0.7821	0.7805
LitLat BERT (litlat-bert)	**0.813**	**0.8126**	**0.8125**	**0.8133**

When comparing the results obtained from the models trained with and without data augmentation, we can observe significant differences. After data augmentation was applied, the performance of LitLat BERT reached a highest accuracy with 0.813 and F1 score of 8 of 0.81.

The improvement in LitLat BERT's performance can be attributed to the fact that it is primarily trained on Lithuanian texts, unlike the other multi-language models. This specialized training allows

LitLat BERT to be more sensitive to the nuances and linguistic patterns specific to the Lithuanian language. When the dataset is augmented, the model can effectively leverage its knowledge of Lithuanian text to adapt to a broader range of linguistic variations, improving its overall performance. Consequently, LitLat BERT could be the preferred choice for emotion identification tasks in Lithuanian due to its targeted training and enhanced adaptability when provided with diverse and augmented data.

6. Conclusions

This study marks the initial effort to translate and adapt a training dataset for downstream NLP tasks in the Lithuanian language. While we did not achieve the same performance level as the original English dataset, our findings demonstrate that models can effectively be trained on translated and augmented data for the Lithuanian context.

In summary, data augmentation, using synonym replacement, is an effective method for enhancing the performance and robustness of emotion identification models in low-resource languages like Lithuanian, by addressing the challenges posed by the morphological richness of the language and generating diverse and balanced training data.

In the future, we intend to accomplish a more in-depth study of how well these services translate in the context of detecting emotions. We also intend to look into ways of improving the translation quality by fine-tuning the translation models.

7. References

[1] Ghafoor, A., Imran, A. S., Daudpota, S. M., Kastrati, Z., Batra, R., & Wani, M. A. (2021). The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. IEEE Access, 9, 124478-124490..

[2] Schuster, S., Gupta, S., Shah, R., & Lewis, M. (2018). Cross-lingual transfer learning for multilingual task oriented dialog. arXiv preprint arXiv:1810.13327..

[3] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196..

[4] Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. Advances in neural information processing systems, 33, 6256-6268..

[5] Bianchi, F., Nozza, D., & Hovy, D. (2022, May). XLM-EMO: Multilingual emotion prediction in social media text. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 195-203)..

[6] Plaza-del-Arco, F. M., Strapparava, C., Lopez, L. A. U., & Martín-Valdivia, M. T. (2020, May). EmoEvent: A multilingual emotion corpus based on different events. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 1492-1498)..

[7] Zygadło, A., Kozłowski, M., & Janicki, A. (2021). Text-Based emotion recognition in English and Polish for therapeutic chatbot. Applied Sciences, 11(21), 10146..

[8] Saravia, E., Liu, H. C. T., Huang, Y. H., Wu, J., & Chen, Y. S. (2018). Carer: Contextualized affect representations for emotion recognition. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 3687-3697)..

[9] Tiedemann, J., & Thottingal, S. (2020, November). OPUS-MT--Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation..

[10] Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... & Joulin, A. (2021). Beyond english-centric multilingual machine translation. The Journal of Machine Learning Research, 22(1), 4839-4886..

[11] "Datasets for Advancing AI Research," META, [Online]. Available: https://ai.facebook.com/tools/.

[12] "DeepL Translate: The world's most accurate translator," DeepL, [Online]. Available: https://www.deepl.com/translator.

[13] "Google Translate," google, [Online]. Available: https://translate.google.com/.

[14] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196..

[15] Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing..

[16] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543)..

[17] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692..

[18] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852..

[19] Ulčar, M., & Robnik-Šikonja, M. (2022, November). Training dataset and dictionary sizes matter in bert models: the case of baltic languages. In Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, Dece.

[20] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805..

[21] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2019). Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781..

[22] Bianchi, F., Nozza, D., & Hovy, D. (2022, May). XLM-EMO: Multilingual emotion prediction in social media text. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 195-203)..

[23] Kim, Joo-Kyung, et al. "Cross-lingual transfer learning for pos tagging without cross-lingual resources." Proceedings of the 2017 conference on empirical methods in natural language processing. 2017..